Expertise and Chess: A Pilot Study Comparing Situation Awareness Methodologies

Francis T. Durso[1], Todd R. Truitt[1], Carla A. Hackworth[1], Jerry M. Crutchfield[1], Danko Nikolic[1], Peter M. Moertl[1], Daryl Ohrt[1], and Carol A. Manning[2]

[1] University of Oklahoma
[2] F.A.A. Civil Aeromedical Institute

Situation awareness (SA) has received a considerable amount of attention in the recent literature. However, no agreed upon definition or methodology for the measurement of this phenomenon currently exists. Many definitions have been proposed and have provided perspectives varied in scope. For example, the definitions provided by Endsley (1988) and Mogford (1994), although not incompatible, present different viewpoints regarding SA. Despite the lack of a common definition, our experiment focused on finding a sensitive procedure that would best differentiate among levels of expertise or skill and hence, SA.

A number of different methodologies have been explored in an effort to understand how operators develop and maintain a "picture" of the situation in which they are involved. Methodologies previously used have included verbal protocol analysis (e.g., Ericsson & Simon, 1994; Ohnemus & Biers, 1993; Sullivan & Blackman, 1991), retrospective event recall (e.g., de Groot, 1965; Kibbe, 1988), concurrent memory probes such as the *Situation Awareness Global Assessment Technique* (SAGAT; Endsley, 1988), and physiological measures such as eye movements (e.g., Moray & Rotenberg, 1989; Stein, 1989; Wierwille & Eggemeier, 1993). Beginning with the assumption that experts have better SA than novices, we compared five very different procedures in one study: verbal protocols, eye movements, on-line queries with the situation present, on-line queries with the situation removed (as in SAGAT), and post-hoc recollection.

We chose to look at chess expertise for several reasons. First, chess has been correctly called the drosophila of cognitive psychology (Charness, 1989), and its long history of study should serve us well in understanding SA. Second, it seems to us that, perhaps more than most activities, differences in chess expertise are differences in SA. The entire game involves assessing the relationship between existing pieces and predicting impending moves. For example, input and output processes are simple and are unlikely to distort the internal model of the situation. Third, chess players are ranked by the United States Chess Federation (USCF) and the differences among rankings are well understood. Finally, our ultimate interest is in understanding SA in air traffic controllers, and chess provides a God's-eye perspective of a number of different entities that make it a nice, albeit limited, laboratory analog of air traffic control.

General Methodology

All participants monitored four chess games. We asked participants to monitor, rather than play, the games to allow control over the entire game. All players experienced exactly the same game regardless of their skill level, allowing us, for example, to insert queries at identical points for each player. To ensure involvement, the participants were asked to monitor the game for imminent material losses. This is a clear component of chess and allowed us to engage the participants even though they were simply monitoring an existing game.

Table 1. Overview of experimental procedures.

	Game 1	Game 2	Game 3	Game 4
Opening	Queen's Gambit (Slav defense/ Main line)	Queen's Indian (Main line)	Caro-Kann (Capablanca /Main line)	Sicilian (Alapin)
Outcome	White mates in 60	Black mates in 74	White mates in 52	Black mates in 61
Number of captures	8	7	12	8
Monitor material loss	Yes	Yes	Yes	Yes
Method tested	Verbal protocols	Eye movements	Situation-present queries	Situation-absent queries (SAGAT like)
Post-hoc recall	Yes	Yes	Yes	Yes

In each game, the participant sat 42" from a projected image of a chess board that subtended a visual angle of 40(, with each square subtending about 5(. At the beginning of each ply, a piece blinked twice, moved, and then flashed twice again. The position remained for 15 sec. The game was stopped after 80 plies (40 moves).

A chess expert (USCF 2100; Expert), an intermediate (USCF 1607; Class B), and a novice (USCF 1374; Class D) monitored four high-level games generated by a commercial computer program, *Chessmaster 4000* (. Characteristics of each game, and what methodological procedures were employed, appear in Table 1. The games were generated by setting both white and black on *Chessmaster's* Chessmaster-level and having the computer play itself. Games were randomly assigned to order of presentation. In all games, moves of both white and black were made while the participant monitored the game to determine when a piece was about to be captured. Across games there were 7 to 12 plies or sequences of plies during which pieces were captured.

In this pilot study, the use of only three participants makes the statistical discovery of the most sensitive procedure problematic. In part we relied on visual inspection of the data, looking for clear differences across levels of expertise. In addition, we conducted item analyses using materials (e.g., queries) rather than participants as the random variate. This allows us to generalize to other probes or situations for these participants, but does not allow us, for example, to discriminate between effects caused by the particular participant and effects caused by his level of skill. All tests were conducted at an alpha level of .05.

Anticipating Material Loss

As the cover task, the participant used a joystick to register his judgment about the imminent loss of material. If he believed that a piece would be taken "in the near future", he was to pull the joystick back; if he believed that a piece would not be taken in the near future, he pushed the stick forward. "Near future" was intentionally left vague so as not to influence the distance in the future that the participant normally considered. Confidence was indicated by the extent to which the stick was pushed or pulled. If the participant did not have any feeling about the upcoming state of affairs, he was to rest the stick in the middle, neutral position.

We began analysis by determining the plies during which a take occurred. To control for strategies (such as always anticipate loss of material) we chose, for each game, a comparable number of plies in which a take did not occur. Thus, if a person did invariably pull the joystick back, he would do well anticipating loss of material, but poorly anticipating when no material loss would occur. Performance of a theoretical perfect player who moved optimally from take to no-take and back was calculated. For each critical ply in the game, we computed the point prior to that occurrence when the participant changed his judgment from "no take" to "take," or from "take" to "no take," and compared it to the theoretical perfect participant. Scores could range from 0 sec, if the ply was never correctly detected, to the theoretical optimum.

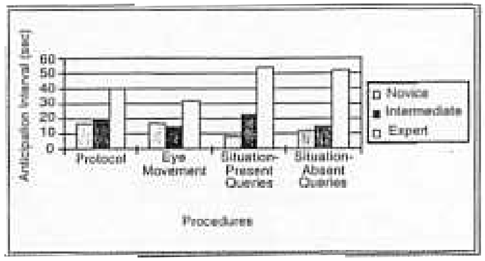

Figure 1. Projecting material loss

Not surprisingly, the expert anticipated material losses (47 sec; 4 plies) sooner than did the class B player (19 sec; 2 plies) or the class D player (13 sec; 1 ply). Results are shown in Figure 1. Regardless of the concurrent procedure, the expert was easily distinguishable from the two nonexperts. A repeated measures ANOVA revealed a significant main effect of skill, *F*(2, 132) = 14.82. Although there was no significant effect of method nor any interaction of method with skill, the figure suggests that only when the participants were in the Situation-present Queries condition were all three levels of expertise distinguishable. This may suggest that the Situation-present Queries condition is most likely to reflect individual differences in SA.

There has been some concern that interrupting the participant could, in itself, interfere with SA (e.g., Sarter & Woods, 1991). However, in the current study, there was little indication for the novice or the intermediate that the particular procedure had any contaminating effect on SA. While verbal protocols have been criticized as obtrusive and unreliable (Nisbett & Wilson, 1977), recording eye movements is touted as a relatively unobtrusive measure of recording behavioral data. However, in our data there was no effect of methodology on anticipating material loss, and what little effect may have been present argues for the use of query techniques.

Verbal Protocols

Verbal protocols were gathered during the first game. Participants were told to "... .talk out loud. Try and verbalize your thoughts about the chess game that you are watching. For example, talk about pieces that are about to be taken, tactics or strategies that you notice are being used, or just general comments about the game such as which side currently has an advantage. We do not expect you to comment about anything in particular, just tell us basically what you are thinking about regarding the game. . . ."

Protocols were transcribed, segmented into plies, and then categorized by the experimenters. Comments relevant to the game were coded into four categories: Predictions, Assessments, Identifications, and Other. Predictions were sentences that described possible future moves (e.g., "The rook will go to H3 and attack the queen."). Assessments were sentences that characterized the ongoing flow of the game without making any predictions (e.g., "White needs to avoid exchanges because he has the developmental advantage."). Identifications were utterances that identified specific moves or tactics (e.g., "The knight is pinned; This is a Queen's gambit."). Comments that did not fit these categories were rare and were excluded from further analyses.

Overall, the intermediate participant made the most utterances (N = 184), the expert the least (N = 94), with our novice (N = 106) falling between these two. This pattern of overall utterances was consistent with the intermediate effect (Grant & Marsden, 1988; Schmidt & Boshuizen, 1993). When the utterances are classified as in Figure 2, it becomes apparent that the participants uttered different types of comments depending on their level of expertise, $\chi^2(4) = 9.49$. Our expert produced mostly predictions (64%), more than either the intermediate or the novice, $\chi^2(2) = 12.46$. The intermediate, on the other hand, produced the largest number of statements assessing the situation, $\chi^2(2) = 76.03$. His corpus was 62% assessment, with only 27% predictions and 10% identifications. Finally, the novice's protocol consisted of 50% assessment, 25% prediction, and 19% identification, a profile too similar to the intermediate's to allow recommendation of this procedure, resulting in a nonsignificant $\chi^2(2) = 4.0$, *NS*.

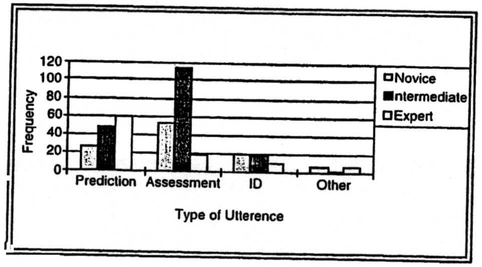

Figure 2. Protocol analysis

Eye Movements

During game two, eye movements were recorded using an Applied Science Laboratories (ASL) series 4000 eye tracker. The eye tracker utilized a magnetic headtracking apparatus in order to compensate for any head movements made by the participant. Eye movement data were recorded in two, 10 minute sessions. The first session included plies 1-40; the second session included plies 41-80. A 10 minute break was allowed between sessions to prevent participants from experiencing any discomfort resulting from the headband of the eye tracker. Participants were instructed to watch the game while using the joystick as in the previous game. Fixations and saccades were recorded. To qualify as a fixation, the eye had to remain looking within .5(visual angle for at least 100 msec.

Unfortunately, in all of our analyses, differences among skill levels were quite small. For example, the largest difference in the fixation rate was 0.4 fix/sec, and the difference in average saccade distance was 1(. A more promising finding may be that the novice player had longer fixation durations (*M* = 283 msec) than did the expert (*M* = 233 msec) or intermediate player (*M* = 198 msec), but even here it was difficult to clearly classify the players. We looked at a myriad of other measures (e.g., fixation duration on critical pieces, area of the board covered) with little evidence that eye movements would be a consistent predictor of expertise. It certainly would not be an easy one to ascertain. Unfortunately, a clear picture of SA, as reflected in eye movement differences, was not apparent, although some interesting trends did emerge. Obviously, this is not to say that such differences do not exist, merely that we could not find them.

Situation-present Queries

During the third game, the participant responded to questions about current and future chess positions. On some trials during the chess game, a tone sounded and a question was presented auditorily while the chess board and pieces remained in view. All participants were asked the same questions on the same plies. Eighteen questions were asked—six from each of three categories: 1) Perceptual, 2) Present Conceptual, 3) Future Conceptual. In this version of the on-line query methodology, the situation remains present while the participant responds. Clearly, the proportion of correct responses should be quite high, given that the participant can determine the correct answer from information still being displayed. Thus, the primary dependent variable was response time.

We looked at this variation of more traditional on-line querying techniques for a number of reasons. First, unlike looking at mistakes in SA, response time allows us to investigate successful SA, rather than inferring characteristics of SA from its failures. Second, our ultimate interest is in exploring SA among air traffic controllers, where any technique that removes them from the radar screen is likely to be disruptive and viewed with suspicion. An auditory query allows the situation to continue and uses an input mode that can be fit into the controller's existing work scheme (e.g., by querying over a telephone line).

The participant answered orally while the experimenter recorded his response. The screen in front of the participant went blank and the response recorded by the experimenter was then displayed, allowing the participant to confirm or change the experimenter's input. A question was asked during 18 randomly selected plies during the game. Participants were queried about: 1) Perceptual characteristics(Where is the white queen?; What piece is adjacent to the black rook?; 2) Present conceptual relations (What piece is the white bishop attacking?; What piece is defending the white knight?; 3) Future relations(What piece can white move to pin black's rook?; What piece can black move to prevent a back rank mate?

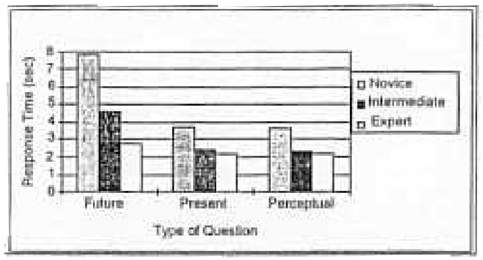

Figure 3. Situation-present queries

As expected, accuracy was quite high and did not distinguish among levels of expertise. Response time, on the other hand, appears to be a good index of expertise, and presumably, SA, *F*(2, 24) = 13.12. Results are shown in Figure 3. Overall, the expert responded faster than the intermediate, who responded faster than the novice. Latency also varied as a function of question type, *F*(2, 12) = 4.78, with future questions taking longer to answer than did the other types. Skill also interacted with question type, *F*(4, 24) = 2.32, *p* <. 10. Differences as a function of skill were present most clearly in questions about future events, *F*(2, 8) = 7.20 and perceptual events, *F*(2, 10) = 5.28.

Situation-absent Queries

More typically, on-line queries freeze the simulation of interest, remove information, and ask the participant a question or series of questions (i.e., Endsley's (1988) SAGAT). In the fourth game, the pieces were removed from the board as a tone sounded, and a question was presented visually to the right of the now empty board. Participants read and answered the question, the screen went blank, the experimenter typed the response to allow the participant to verify the entry, and then the board reappeared as it was when the question sequence was initiated. As before, 18 questions were asked.

The data from situation-absent procedures are the number of questions responded to correctly. Those data appear in Figure 4. It is worth noting that the figure presents results similar to those found in the verbal protocol: an expert advantage for predictions (future) and an intermediate advantage for assessments of the (current) position. This methodology showed a marginal skill effect, *F*(2, 30) = 2.94, *p* < .10. Questions about the future distinguished most clearly among the three skill levels, *F*(2, 10) = 3.18, *p* < .10. Skill differences were not reliable for the perceptual or present questions, despite the graph's suggestion of an intermediate superiority for present-queries.

Given the success we had with the response time analysis in the situation-present procedure, we looked at the response times associated with the correct responses in the situation-absent procedure. Unlike percent correct, visual inspection of the response time data revealed consistent differences among skill levels (although not as clear as the situation-present condition). Unfortunately, the error rate prevented any meaningful analysis of correct latency. Nevertheless, this may suggest that the differences between the situation-present and situation-absent procedures may be relatively unimportant, provided that response time is used to assess differences among levels of expertise.

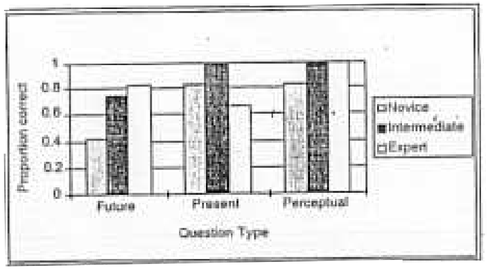

Figure 4. Situation-absent queries

Memory

The final analysis involved consideration of the recall protocols participants gave after the game. Although recall occurred after each game, we considered only the first game here, since it would be uncontaminated by other recalls. The participants differed dramatically in the length of their recall protocols, with the expert saying very little and the intermediate saying quite a bit. However, the expert's recall consisted primarily of general abstractions that characterized large segments of the game. The novice's recall consisted of several move by move recollections of the game. The intermediate's recall fell between these two. Unfortunately, the succinctness of the expert's recall made it difficult to analyze the protocols beyond this overall classification. However, it does suggest that experts tend to convey labels of encapsulated information (Schmidt & Boshuizen, 1993).

Discussion

We investigated five procedures. Of those, eye movements seemed to be the most complicated and yielded the fewest insights. Analysis of memory protocols and on-line protocols were also problematic. The two query procedures seemed to supply some useful information. Across the procedures, we found considerable evidence that questions about the future are most likely to discriminate among all three levels of expertise. Number of predictions in the verbal protocols, proportion correct for future-oriented queries (situation-absent), and response time for future-oriented queries (situation-present) all suggest that researchers interested in distinguishing levels of SA would do well to focus their efforts on future events. Even our cover task, which required participants to predict material loss, showed clear expertise effects. Information about the current state of affairs, as measured by assessment utterances in verbal protocols and proportion correct in